Agriculture, Ecosystems and Environment 254 (2018) 213–223

Contents lists available at ScienceDirect

Agriculture, Ecosystems and Environment

journal homepage: www.elsevier.com/locate/agee

Baseline map of organic carbon stock in farmland topsoil in East China

Xunfei Deng[a,b], Xiaojia Chen[a,b], Wanzhu Ma[a,b], Zhouqiao Ren[a,b], Minghua Zhang[c], Michael L. Grieneisen[c], Wenli Long[a], Zhihua Ni[d], Yu Zhan[e,*], Xiaonan Lv[a,b,**]

[a] Institute of Digital Agriculture, Zhejiang Academy of Agricultural Sciences, Hangzhou, Zhejiang 310021, China
[b] Key Laboratory of Information Traceability for Agricultural Products, Ministry of Agriculture, Hangzhou, Zhejiang 310021, China
[c] Department of Land, Air, and Water Resources, University of California, Davis, CA 95616, USA
[d] Agricultural Technology Extension Service Center of Zhejiang Province, Hangzhou, Zhejiang 310029, China
[e] Department of Environmental Science and Engineering, Sichuan University, Chengdu, Sichuan 610065, China

ARTICLE INFO

Keywords:
Soil organic carbon
Baseline map
Machine learning
Carbon accounting
Harmonized world soil database
Climate change

ABSTRACT

Soil organic carbon (SOC) is important to soil fertility and the global carbon cycle. Accurate estimates of SOC stock and its dynamics are critical for managing agricultural ecosystems and carbon accounting under climate change, especially for highly cultivated regions. We extensively surveyed the SOC levels in 29,927 sites in Zhejiang province, an intensively cultivated region of East China, from year 2007 to 2008. We then estimated the spatial distribution of topsoil (0–30 cm) organic carbon stock using a random forest (RF) model, which is a powerful machine learning algorithm with superior predictive performance over parametric statistical models. The final RF model contained 23 predictor variables, covering soil properties, vegetation, climate, topography, land cover, farming practices, and locations. The RF model showed high performance in predicting the SOC stock, with a coefficient of determination (R^2) of 0.76 and a root mean square error (RMSE) of 10.63 t C ha^{-1}. This performance was superior to the General Linear Model (GLM) (R^2 = 0.35, RMSE = 19.93 t C ha^{-1}) and the ordinary kriging (OK) method (R^2 = 0.57, RMSE = 14.44 t C ha^{-1}), and was equivalent to Boosted Regressing Trees (BRT) (R^2 = 0.73, RMSE = 11.26 t C ha^{-1}). According to the variable importance evaluation, soil properties were the most important predictor variables, followed by climate and location, with relative importance values of 61%, 17%, and 14%, respectively. The predicted SOC stock ranged from 14.8 to 125.5 t C ha^{-1}, with an average ± standard deviation of 50.1 ± 12.3 t C ha^{-1}. The mean SOC level obtained from this survey was considerably lower than the value of 60.5 t C ha^{-1} reported for the same region in the Harmonized World Soil Database (HWSD), which is the most commonly used soil database worldwide. A large spatial discrepancy of SOC stock was observed between this survey and HWSD in regional and sub-regional levels. This study provided an updated regional baseline map of SOC levels for improving farmland management and refining carbon accounting under climate change.

1. Introduction

There are increasing concerns over intensified carbon dioxide (CO_2) emissions and climate change (Bell and Worrall, 2009). Soil is an important carbon pool in the global carbon cycle since it stores the most terrestrial organic carbon (Lal, 2008). Soil carbon sequestration mitigates the greenhouse gas effect and improves soil fertility (Viscarra Rossel et al., 2014). Recently, increased attention has been paid to the SOC dynamics, particularly with disturbances caused by anthropogenic activities under climate change (Pan et al., 2010; Muñoz-Rojas et al., 2015). Around the world, baseline maps of soil organic carbon (SOC) are extremely important for understanding the soil carbon balances on regional and global scales, and these maps have become indispensable components of Intergovernmental Panel on Climate Change (IPCC) Guidelines for Greenhouse Gas Inventories (IPCC, 2006).

SOC stocks and their spatial distributions have been reported on various spatial scales in the past. At national or continental levels, the SOC stocks were estimated in Europe (Lugato et al., 2014), Africa (Hengl et al., 2015), North America (Mishra et al., 2010), and China (Ni, 2001). Most of these SOC baseline maps were derived from limited soil survey data collected in previous decades, resulting in coarse spatial resolutions and a tendency to be very out of date. Globally, the Harmonized World Soil Database (HWSD) with the world soil map (FAO-90) has aimed to meet the requirements for soil carbon

* Corresponding author.
** Corresponding author at: Institute of Digital Agriculture, Zhejiang Academy of Agricultural Sciences, Hangzhou, Zhejiang 310021, China.
E-mail addresses: yzhan0@gmail.com (Y. Zhan), luxn@mail.zaas.ac.cn (X. Lv).

https://doi.org/10.1016/j.agee.2017.11.022
Received 29 June 2017; Received in revised form 22 November 2017; Accepted 23 November 2017
Available online 22 December 2017
0167-8809/ © 2017 Elsevier B.V. All rights reserved.

measurements under the Kyoto Protocol and the FAO/IIASA global agro-ecological assessment study. It is the most commonly used SOC baseline in the world, and has a spatial resolution of approximately 1 km (FAO/IIASA/ISRIC/ISSCAS/JRC, 2012). HWSD has become the main data source for estimating large-scale SOC storage and is currently widely used by climate change models (Jobbagy and Jackson, 2000; Batjes, 2009; Perlman et al., 2014). These large-scale SOC baselines provide valuable bases for carbon accounting as well as modeling interactions between the global carbon cycle and climate change (Liu et al., 2006; Mishra et al., 2010; Romero et al., 2012).

While SOC levels generally remain steady for decades, dramatic land use changes and intensive agricultural activities lead to considerable changes of local SOC. This results in high discrepancies between the current situation and the SOC levels recorded in HWSD, as the data in that source were derived from the soil surveys conducted decades ago (FAO/IIASA/ISRIC/ISSCAS/JRC, 2012). SOC contents are continuously influenced by environmental conditions and land use changes (Fantappiè et al., 2010; Martin et al., 2010), especially for intensively cultivated farmlands (Six et al., 2000; Halvorson et al., 2002; Yan et al., 2012). SOC turnover might also be induced by shifts in soil management practices (West and Post, 2002; Hengl et al., 2015). For instance, increases of regional SOC stock were found in China and the USA due to enhanced agricultural production or land use changes (Liao et al., 2009; Pan et al., 2010; Xiong et al., 2014). Furthermore, the quality of the SOC baseline map is highly affected by sampling densities. Data interpolation or extrapolation from a limited number of soil samples on the basis of soil map units often overlook the effects of climate, land cover, and soil management practices (Smith et al., 2012). In addition, the uncertainties associated with the SOC baselines tend to be propagated in downstream analyses, such as carbon accounting for the global carbon cycle or modeling greenhouse gas emissions (e.g., Batjes, 2009; Perlman et al., 2014; Ma et al., 2016).

Many studies have derived accurate spatial SOC stock on regional or national scales using various approaches (Mishra et al., 2010; Elbasiouny et al., 2014). Calculating SOC averages by soil types and spatial interpolation from site-specific observations have been commonly used (Batjes, 2009; Chaplot et al., 2010). In addition, regression models, such as general linear model (GLM) with a set of predictor variables (e.g., land uses, soil types, and management practices), were developed to capture the SOC spatial variation (Mishra et al., 2010; Meersmans et al., 2012). Moreover, various machine leaning algorithms (e.g., random forest and boosted regression trees) have been used to develop SOC baselines at fine spatial resolutions based on existing soil databases and auxiliary data, such as satellite retrievals and climatic conditions (Viscarra Rossel et al., 2014; Were et al., 2015; Hengl et al., 2017; Schillaci et al., 2017). These regional estimates of SOC stock tend to be more accurate than the estimates solely derived from the site observations, as the auxiliary information carried in the predictor variables improves the interpolation accuracies. The estimates at higher resolutions (e.g., ~250 m or ~90 m) have greatly facilitated sophisticated regional simulations (Lugato et al., 2014; Hengl et al., 2015). However, the large-scale assessment of baseline SOC stock at fine spatial resolutions are typically unavailable due to the lack of extensive soil survey data and the sophisticated statistical modeling that is required (Viscarra Rossel and Chen, 2011).

East China is one of the most intensively cultivated regions in the world. The most recent national soil survey was conducted in early 1980s, and while it is the data source for HWSD it is woefully inadequate to represent the current status of SOC in this region (Pan et al., 2010). It is therefore important to update the SOC baseline map. This study mapped the spatial distribution of SOC stock (0–30 cm) in the farmlands of Zhejiang province, East China. Random forest is an ensemble of classification or regression trees which has shown higher predictive performance than other statistical models in various environmental and ecological research fields (Cutler et al., 2007; Reid et al., 2015). Recently, random forest has been employed to estimate various soil properties in digital soil mapping (DSM) (Grimm et al., 2008; Were et al., 2015; Hengl et al., 2017; Zhang et al., 2017). In our region, a random forest model was developed with a comprehensive set of predictor variables, including vegetation, topography, climate, land use, locations, farming practices, and soil properties. The regional SOC stock distributions were estimated at a fine spatial resolution (~250 m), and estimation uncertainties were also assessed. On the basis of the estimated SOC distributions, we then evaluated the differences in the SOC stock distributions between this study and the HWSD. The high-resolution SOC baselines mapped in this study are critical for improving regional farmland management by promoting good practices associated with elevated SOC stock, and refining carbon accounting under climate change with the updated SOC stock data.

2. Materials and methods

2.1. Study area

Zhejiang province (27°02′N-31°12′N, 118°01′E-122°58′E; ~102,646 km^2) is one of the most highly developed regions in East China. The primary land covers are forest and farmland. The farmlands are mainly located in the Hangjiahu plain, central basin, and east coast region, where the elevations are generally lower than 100 m (Fig. S1). Paddy rice, vegetables, dry-land crops (e.g., broad bean, maize, and oil rape) and their rotations are the dominant cropping patterns in these agricultural areas. The main soil types are Fluvisols, Anthrosols, Acrisols, Alisols, and Regosols (ZJSSO, 1994). This region is located in the subtropical climate zone, with an annual mean temperature of 15–18 °C and an annual mean precipitation of 980–2000 mm. The precipitation and temperature exhibit increasing gradients from north to south, with local variations occurring based on topography (CMDC, 2015).

2.2. Field survey and laboratory analysis

A total of 29,927 topsoil (0–20 cm) samples were collected from May 2007 to April 2008 by employing the standard soil sampling protocol (Fig. 1) (Wang and Zhou, 2011). Note that the 20cm-depth SOC measurements were used to estimate the 30 cm-depth SOC by using the method employed in a previous study (Yu et al., 2012) in order to compare our data with data in the HWSD. Please refer to S1 in the Supplementary material for the details. The digital maps of soil types, cropping patterns, and land use were spatially overlaid, and their intersected polygons were identified as the spatial units for designing the field survey. This spatial-unit map was used to determine the locations of sampling sites in order to best capture the spatial variations in SOC. In the field survey, one soil sample was collected per 0.25–0.4 km^2. In farmlands with frequent crop rotations, resulting in intense SOC turnover, the sampling density was increased to approximately one soil sample per 0.15–0.25 km^2. Each sample consisted of 6–8 soil sub-samples, which were collected by following an “S” shape at intervals of a few meters. All of the sub-samples were collected at a depth of 20 cm and had a similar weight of approximately 0.5 kg. The methods of coning and quartering were employed to adequately mix these sub-samples in order to obtain one representative soil sample per site with the weight of 1 kg. Note that soil sampling was conducted at least a week after the previous rotation crop was harvested in order to exclude the effects of fertilization.

During the field survey, the previous rotation crop, topography of the farmland, and the longitude/latitude of each representative soil sample were also recorded. The soil texture of the sampling site was obtained from the field estimation. A commonly used method for estimating soil bulk density was conducted with five parallel measurements at each sampling site. To collect a fixed volume (100 cm^3) of topsoil, a metal ring with an intact core was pressed into the soil, followed by careful excavation of the soil around the ring without disturbing or

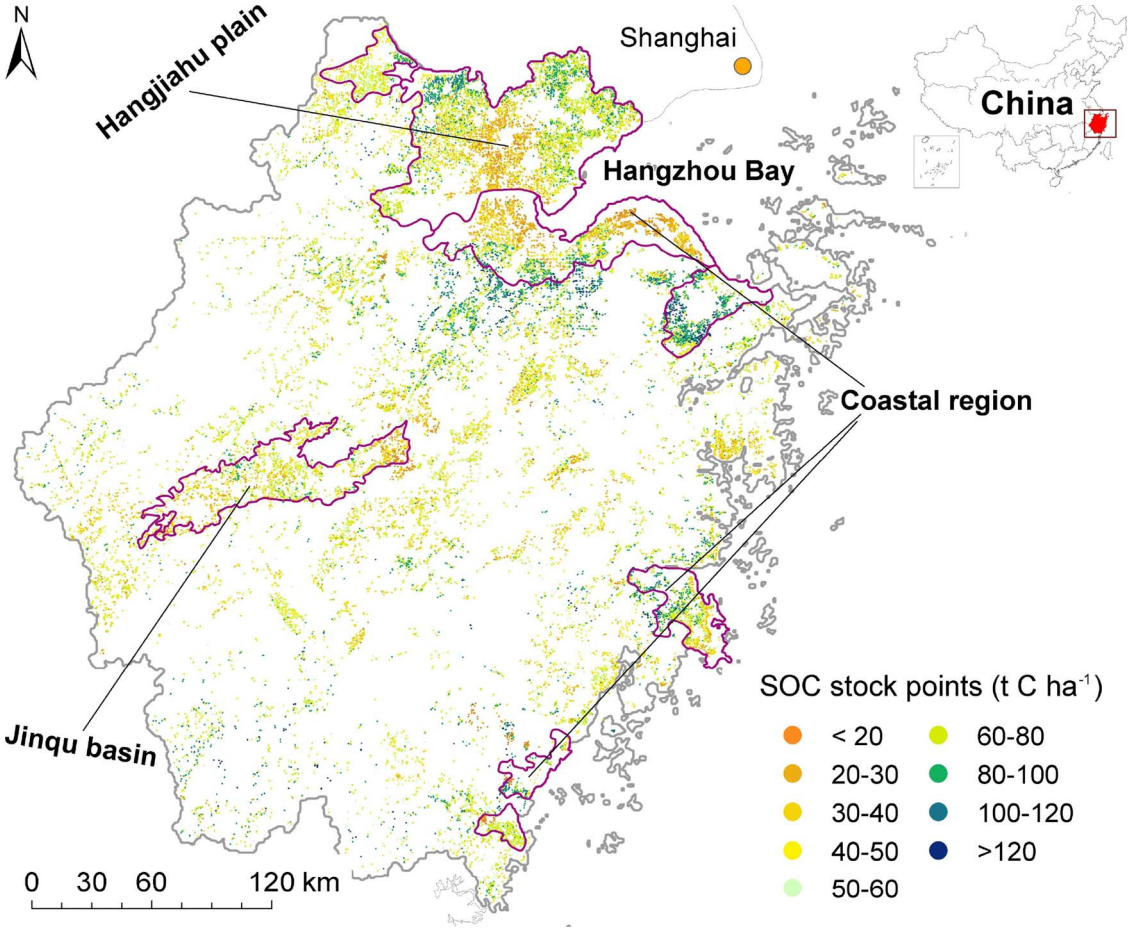

Fig. 1. Spatial distributions of the soil sampling sites for farmland soil in the study area. A total of 29,927 topsoil (0–20 cm) observations were collected from April 2007 to May 2008. On the basis of the 20 cm-depth measurements, the 30 cm-depth SOC stock was estimated at each sampling site to be comparable with Harmonized World Soil Database (HWSD). Please refer to S1 in supplement material for estimation equation.

loosening the soil contained in the ring. The remaining soil contained in the ring was then poured into a plastic bag for further laboratory analysis. The soil bulk density was estimated by dividing the dry weight (dried at 105 °C) by the volume of soil. The soil samples for all other analyses (except for soil moisture) were air-dried at ambient temperatures, and root residues in the soil were removed. Each soil sample was ground to pass through a 2 mm sieve for further analysis of soil properties, such as pH, total nitrogen, and available phosphorus (Table 1). The remaining soil (fragments > 2 mm) was weighed to obtain the coarse fragment content. Approximately 50 g of ground soil was collected using the coning and quartering method, and further grinded to completely pass through a 0.25 mm sieve for the SOC content analysis. The SOC content was measured using the potassium dichromate oxidation method based on the changes of dichromate ion ($Cr_2O_7^{2-}$) after the oxidation of soil organic carbon in the heated mixed solution (MOA, 2006). Triplicate SOC measurements were performed for each soil sample.

2.3. Data sources and preparation

The complete set of predictor variables (shown in Tables and S2) were derived from meteorological conditions, vegetation, topography, land cover, location, farming practices, and soil properties. The meteorological data from May 2007 to April 2008, including evaporation, ground surface temperature, air temperature, precipitation, air pressure, and sunshine, were observed at meteorological stations and published by the China Meteorological Data Center (CMDC, 2015). The Normalized Difference Vegetation Index (NDVI) data at 250 m spatial resolution were retrieved from the Moderate Resolution Imaging Spectroradiometer (MODIS: MOD13Q1 C006, Tiles: h28v05 and h28v06) (NASA LP DAAC, 2016). The elevation data at a spatial resolution of 30 m were downloaded from the NASA Shuttle Radar Topography Mission (SRTM) 1 Arc-Second Global (2000) (NASA LP DAAC, 2015). The land use classifications were derived from the Global Land Cover 30 m (Globalland30, 2010) product distributed by the National Administration of Surveying, Mapping and Geoinformation of China (NASMG, 2016). A digital atlas of Zhejiang for year 2008, which includes farmlands, villages, coastlines and roads, was provided by the Zhejiang Administration of Surveying, Mapping and Geoinformation (ZASMG, 2013). In addition, all the soil parameter predictors obtained from the field survey were processed to a vector (point) format for subsequent spatial interpolation.

A 250 m × 250 m grid was defined for the SOC mapping. All of the input data were resampled to this grid for the model development and prediction. The meteorological conditions (e.g., temperature and precipitation) for each cell were interpolated from 17 meteorological sites enclosing the study area through co-kriging with elevation (Deutsch and Journel, 1998). To improve the accuracy of predictor variable data generated through spatial interpolation at both the training and prediction stages, the site-based observation data (e.g., meteorological conditions) were transformed to normal distributions by using the normal score transformation (Deutsch and Journel, 1998), and then interpolated to the grid cells by using ordinary kriging implemented in the autoKrige program, where the best variogram model among the Spherical, Gaussian, Matern, and Exponential models was selected for spatial interpolation (Hiemstra et al., 2009). The interpolated values

Table 1
Predictor variables for modeling soil organic carbon stock in farmland.

Variables category	Variables	Data range
Vegetation	Normalized Difference Vegetation Index	−0.3 to 1.0
	Previous crop before the soil sampling	27 classifications, e.g., paddy rice, vegetables, oil rape, broad bean.
Land cover	% of cultivated land within 2.5 km buffer	0–100
	% of forest within 2.5 km buffer	0–100
	% of water bodies within 2.5 km buffer	0–99.2
	% of artificial surface within 2.5 km buffer	0–100
Climate	Mean evaporation (mm)	2.05–3.45
	Mean ground surface temperature (°C)	13.69–24.96
	Mean air temperature (°C)	17.89–22.62
	Mean annual precipitation (mm)	892.3–1743.1
	Mean air pressure (hPa)	995.4–1019.6
	Mean sunshine duration (h day^{-1})	3.95–5.64
	Mean relative humidity (%)	63.8–79.2
Soil properties	Soil texture	Clay, Clay loam, Loam, and Sandy soil
	Soil type	Fluvisols, Anthrosols, Acrisols, Alisols, and Regosols
	Percentage of soil coarse fragment (> 2 mm) (%)	0–16
	Water table depth (cm)	40–180
	Soil moisture (%)	2.3–92.1
	Soil salinity (g kg^{-1})	0–12.1
	Soil pH	3.6–9.4
	Cation exchange capacity (cmol kg^{-1})	1.7–98.3
	Soil total nitrogen (g kg^{-1})	0.22–3.18
	Available phosphorus (mg kg^{-1})	0.86–423.9
	Available potassium (mg kg^{-1})	1.0–539.2
Topography	Farmland slope (degree)	0–25
	Farmland topography classes	Valley plain, alluvial plain, coastal plain, mountain terrain
	Elevation (m)	0–1314
Geographical location	Distance to sea (km)	0.1–242.1
	Distance to village (km)	0.04–18.82
	Longitude (°)	118.03–122.39
	Latitude (°)	27.18–31.15
Farming practices	Averaged nitrogen-fertilizer application rate (kg ha^{-1})	68.5–356.3

were then back-transformed to their original scales. In addition, the elevation for each cell was calculated as an area-weighted average through spatial overlay. The maximum value compositing (MVC) method was adopted to obtain the NDVI value for each cell (Holben, 1986). For land use, 2.5 km-radius buffer zones were created to obtain surrounding land cover compositions (i.e., cultivated land, forest, water bodies, and artificial surface) for each cell. The distances from coast (or village) were defined as the shortest distance from the coastline (or village) to the geometric center of each cell. In addition, the categorical soil properties (e.g., soil texture) and county-level nitrogen-fertilizer application rates were assigned to the grid cells through spatial overlay.

2.4. Variable selection and model validation

The complete set of predictor variables for the random forest model was derived from field survey and public data sources (Tables and S2). It has been well-established that the factors of topography, vegetation, climatic conditions, farming practices, and soil properties affect the SOC variability to various extents (Jobbagy and Jackson, 2000; Miller et al., 2004; Kemmitt et al., 2006; Fantappiè et al., 2010). Therefore the random forest model used here initially included the slope, elevation, farmland topography, NDVI, growing crops, land cover, meteorological data (e.g., temperature, precipitation, and sunshine duration), fertilizer application rate, soil total nitrogen, soil type, soil texture, and soil moisture. As the soil bulk density was used in the equation of SOC stock estimates (Eq. (S1)), it was excluded from the random forest model. In addition, the other predictor variables such as the distance to village and the soil salinity were considered to be related to SOC formation, as organic manure tended to be used near villages and the high soil salinity constrained biomass production (Li et al., 2014).

Additionally, the random forest model is capable of simultaneously handling categorical and continuous variables, as well as complex variable relationships such as nonlinearity and interactions (Breiman, 2001). While the random forest model is relatively robust with respect to collinearity among predictor variables, the high-correlated or redundant predictor variables tend to increase the model estimation uncertainty (Dormann et al., 2013). Thus, backward variable selection was conducted on the random forest model, where the predictive performance at each step was evaluated using 10-fold cross-validation, and the least important variables were removed from the model one at a time. The model with the variable subset that exhibited the best predictive performance in terms of lowest RMSE and highest R^2 in the cross-validations was selected as the final model. In the present study, the random forest model with 23 selected predictor variables was chosen to estimate the SOC stock in Zhejiang province. The resulting hyper-parameters of the random forest model were tuned for optimal predictive performance, where *mtry* (i.e., number of variables randomly sampled as candidates at each split) varied from 1 to 23 and *ntree* (i.e., number of trees to grow) varied from 500 to 2000. Finally, an optimized ensemble of regression trees and their parameters was constructed at the training stage and then for the model prediction.

Furthermore, the predictive performance of the random forest model was evaluated by calculating the coefficient of determination (R^2) and root mean square error (RMSE) against a “holdout” dataset. Approximately 10% of data in the overall dataset (n = 3000) were randomly selected for validation, and the other 90% were used for model training. This validation strategy was repeated three times to obtain an averaged R^2 and RMSE for the random forest model, as well as comparison to other models, including GLM, Ordinary Kriging (OK), and Boosted Regression Tree (BRT) (Elith et al., 2008). The importance of each predictor was determined by measuring increases of prediction error when the out-of-bag (OOB) sample for that predictor was

permuted while all other predictors remain unchanged (Liaw and Wiener, 2002). The partial dependence plot, which illustrates the effect of a predictor variable on the SOC level after considering the mean effects of all the other predictor variables (Elith et al., 2008), was employed to evaluate the relationships between individual predictor variables and the SOC level.

2.5. Uncertainty analyses

As the data of several predictor variables were obtained from spatial interpolation, the sequential Gaussian geostatistical simulation (SGGS) was employed to evaluate the prediction uncertainties that originated from these input data (Deutsch and Journel, 1998). SGGS was used to reconstruct realistic distributions, rather than working with the "best" but over-smoothed estimate for a particular site obtained by kriging. In total 100 SGGS simulation samples were generated for the uncertainty analysis. Then, the prediction uncertainties, expressed as 95% confidence intervals (CI), that emerged from the input data uncertainties were estimated for each grid cell, regions of interest, and the whole study area. Furthermore, the coefficients of variation (CV) for SOC predictions were estimated from the variance of bagged predictors using the bias corrected versions of infinitesimal jackknife estimators (Wager et al., 2014).

2.6. Statistical analyses

The statistical analyses for variable correlations, importance measurements, SOC predictions, and satellite retrieval processing (i.e., NDVI, elevation, and land cover), were performed on the R computing platform (R Core Team, 2016). ArcGIS 10.2 was used for spatial data processing and mapping.

3. Results and discussion

3.1. Descriptive statistics of sampled SOC stock

The sampled SOC stock of farmland topsoil (0–30 cm) generally followed a Gaussian distribution, varying from 1.18 to 213.55 t C ha^{-1}, with mean ± standard deviation (μ ± σ) of 49.74 ± 20.31 t C ha^{-1}. The first, second, and third quartiles were 34.61, 48.84, and 63.10 t C ha^{-1}, respectively. The sampled SOC stock showed a high spatial heterogeneity across the study area with the fitted parameters range, nugget, and partial sill of 56,953 m, 582, and 861, respectively (Fig. S2). The nugget effect reflected the variability of SOC measurements at distances shorter than the typical sampling intervals (~500 m) in this study, and included the uncertainty of the sampled SOC data. The high SOC stock was observed in the hinterland of southern Hangzhou bay and the outage of Hangjiahu plain, while the low SOC stock was found in the central Hangjiahu plain, coastal region, and basin or valley terraces in middle-west (Fig. 1). The average SOC stock (μ ± σ) in the Hangjiahu plain, coastal region, and Jinqu basin were 50.49 ± 19.69, 47.69 ± 23.55, and 38.61 ± 15.87 t C ha^{-1}, respectively.

3.2. Baseline of organic carbon stock

High spatial heterogeneity was found on the SOC baseline from the farmland of Zhejiang. The overall organic carbon stock in farmland topsoil (0–30 cm) was estimated to be 50.1 ± 12.3 t C ha^{-1} (μ ± σ), ranging from 14.8 to 125.5 t C ha^{-1}. The total SOC stock in Zhejiang was 90.6 ± 24.5 Mt C (μ ± σ) (Table 2). The spatial distributions of SOC stock were mapped at a 250 m × 250 m resolution and (Fig. 2a). Farmlands located around Hangzhou Bay (e.g., central Hangjiahu plain and the northern coastal plain) and in central areas (e.g., Jinqu basin) had relatively lower SOC stock of less than 40 t C ha^{-1} (Fig. 2a; Table 2). Higher SOC stock was found in the hinterland of southern Hangzhou Bay, the outage of Hangjiahu plain, and the valley terraces (Figs. 1 and 2a). The estimates of the SOC stock in this study were similar to the previous studies (Table S1), but lower than that recorded in HWSD (Table 2).

Recent studies for estimating recent organic carbon in agricultural soils have generally used process-based models, such as DeNitrification-DeComposition (DNDC) (Tang et al., 2010), Agro-C (Yu et al., 2012), and RothC (Farina et al., 2017). Most of these studies, however, focused on the temporal SOC changes without considering the spatial distributions, and were therefore not suitable for spatially resolved environmental modeling (Romero et al., 2012). A recent spatial SOC baseline map, which is the most spatially explicit SOC stock for agricultural soils at large scales thus far, was developed for Europe based on the available soil databases and the CENTURY model simulation (Lugato et al., 2014). The initial input SOC values in these studies were generally derived from the soil inventories of previous decades and the relatively sparse data that were available were used for model calibration and evaluation. However, spatially extensive soil sampling data generated by this study is indispensable for running these environmental models (Batjes, 2009; Romero et al., 2012; Stoorvogel et al., 2017).

The newly estimated regional SOC stock in this study is the most recent and spatially resolved SOC baseline map (~250 m resolution) for Zhejiang province. The previously reported baselines of regional SOC stock for this study area were based on the averages over the soil map units (Table S1), which were derived from the very sparse spatial resolution soil inventories of the second national soil survey during 1978–1984 (ZJSSO, 1994). The SOC stocks in these previous studies are similar to those recorded in HWSD. In China, a recent SOC baseline map for general soils was produced at a 2 km spatial resolution for Jiangsu province (Liao et al., 2009). Brief statistical results of SOC content were reported at the national scale based on field surveys conducted from the early 1980s to 2007, but a spatially resolved baseline map was not provided (Pan et al., 2010). Globally, two newly reported SOC baselines for Australia and Africa, using 5,588 and 28,000 sampling sites, respectively, are the only updated large-scale baselines at fine spatial resolutions (90 m and 250 m) (Viscarra Rossel et al., 2014; Hengl et al., 2015). Our new spatially explicit SOC baseline can support a more detailed regional carbon accounting under climate change using the IPCC Tier 3 (T3) method, and include more sophisticated models and inventory measurements, to provide more accurate results (IPCC, 2006; Villarino et al., 2014).

3.3. Predictive performance

The final random forest model with the 23 selected predictor variables showed a high predictive performance, with R^2 of 0.76 and RMSE of 10.63 t C ha^{-1} (Fig. 3). For the predictive performance, the random forest model is superior to the GLM (R^2 = 0.35, RMSE = 19.93 t C ha^{-1}) and the ordinary kriging method (R^2 = 0.57, RMSE = 14.44 t C ha^{-1}), and is equivalent to BRT (R^2 = 0.73, RMSE = 11.26 t C ha^{-1}) (Fig. S3). Generally, random forests have consistently shown superiority over the generalized regression models in predicting soil and other ecological environment elements (Cutler et al., 2007; Zhang et al., 2017). While kriging has been widely used to obtain the predictions for unsampled locations, its main weakness is that the spatial predictions tend to be over smoothed (Olea and Pawlowsky, 1996). Also, ordinary kriging cannot utilize the auxiliary elements related to SOC stock, and previous studies have found lower predictive accuracy of kriging compared to random forest (Li et al., 2011). In addition, the random forest model also showed superior predictive performance for estimating regional SOC/SOM content over studies using BRT and Support Vector Regression (SVR) models (Martin et al., 2014; Were et al., 2015). The performance of this study was similar to a national SOC stock study in Australia using a data mining algorithm CUBIST, which is a form piecewise linear decision tree model

Table 2
Estimates of soil organic carbon stock and their standard deviations for the whole study area and regions of interest, as well as the comparisons to the Harmonized World Soil Database (HWSD).

Regions[a]	Farmland area[b] ($\times 10^4$ ha)	This Study SOC Stock (t C ha^{-1})	Total SOC Stock (Mt C)	HWSD SOC Stock[c] (t C ha^{-1})	SOC stock difference[d] (%)
Jinqu basin	15.2	41.1 ± 7.4	6.3 ± 1.1	61.6 ± 26.9	−33.3
Hangjiahu plain	32.5	51.3 ± 12.7	16.7 ± 4.1	48.5 ± 18.4	5.8
Coastal plain	26.0	48.8 ± 18.8	12.7 ± 4.9	40.6 ± 18.2	20.4
Valley terrace[a]	17.6	55.2 ± 10.5	9.7 ± 1.8	75.4 ± 33.2	−33.4
Study area	198.7	50.1 ± 12.3	90.6 ± 24.5	60.5 ± 29.7	−17.1

[a] The valley terrace is a cell where more than 75% of land cover within 2.5 km buffer is forest, and the other regions are shown in Fig. 1.
[b] The areas were estimated according to the total cultivated land data of 2009 revealed by the Department of Land and Resources (MLR, 2014).
[c] The SOC stock was derived from the Harmonized World Soil Database (HWSD) database, and the HWSD grids within each region that overlapped our farmland cells were selected to compute mean values and standard deviations.
[d] The percentage of regional SOC stock difference between HWSD and our study was expressed as the mean difference divided by the mean of HWSD, and the negative values indicate the SOC loss.

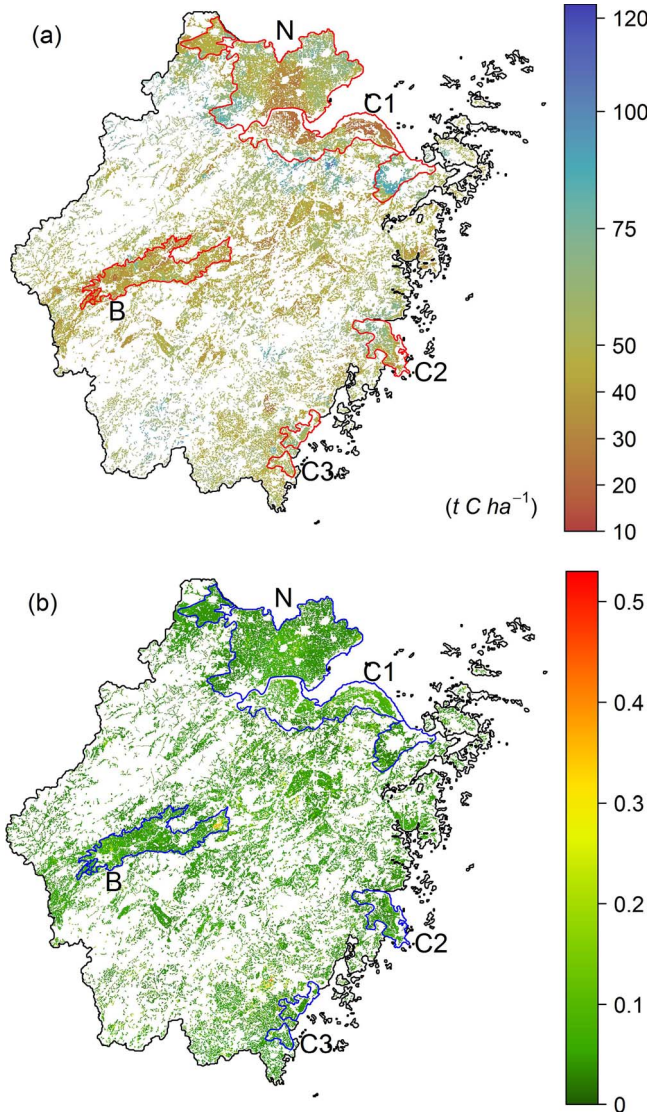

Fig. 2. (a) Estimates and (b) uncertainties of the SOC stock for farmlands in Zhejiang province. Uncertainties are indicated by coefficients of variation (CV), expressed as the standard deviation divided by its mean. Sub-regions of interest are labelled on the maps: Hangjiahu plain (N), coastal plain (C1, C2, C3), and Jinqu basin (B). The areas rendered with white color in (a) and (b) are non-farmlands.

(R^2: 0.75–0.82) (Viscarra Rossel et al., 2014). These different performances might result from either the different algorithms and/or predictor variables used, or from the spatial scale variation and/or regional

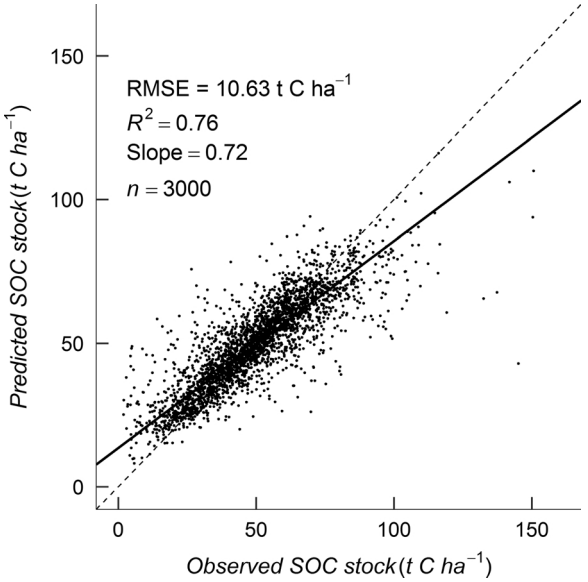

Fig. 3. The model performance of the final random forest (RF) model with the 23 selected predictor variables. The model validation was based on an independent dataset (n = 3000) randomly selected from complete dataset (29,927 soil observations). The R^2, RMSE, and slope were averaged from three randomly chosen validation datasets, and the dataset closest to the average performance was used in charting. The dashed line is the 1:1 line.

characteristics.

The good predictive performance of the random forest model has been ascribed to its data-adaptive capability in handling interactions and nonlinear relationships (Breiman, 2001). High correlations (e.g., elevation versus climatic conditions) and nonlinearity (e.g., SOC versus soil pH) have commonly existed in the predictor variables (Figs. S7 and S8). These inter-correlations and nonlinearity could not be handled well by generalized regression models using parameterizing model structures (Meersmans et al., 2012), and the results can suffer from substantial precision losses if the variables are not in right interaction terms. Instead, tree related methods such as random forest are quite resilient to correlated features (i.e., input predictor variables) and collinearity for random feature selection (Gromski et al., 2014). In our model training with many input features, if one feature is chosen to make a split, the other correlated feature will not be used below the split, as it would never reduce impurity. Furthermore, the redundant variables excluded in model optimization would likely benefit for the improvements of model prediction (Fig. S4).

3.4. Uncertainty analyses

The input data uncertainty is a trivial factor for achieving reliable

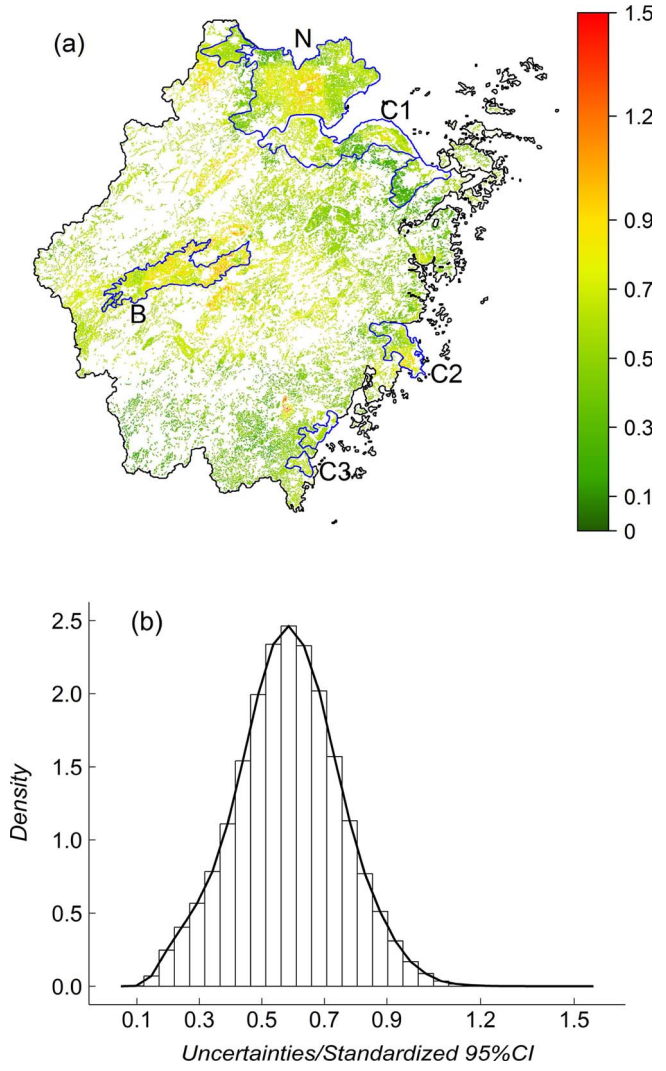

Fig. 4. (a) Map of data-induced uncertainties and (b) their probability distribution for SOC stock in farmlands from model input simulations expressed in standardized form as the range of the 95% confidence intervals (CI) divided by its mean. The standardized 95% CIs of SOC stock prediction caused by input data uncertainty were 35.40–63.21 t C ha^{-1}, 35.54–66.40 t C ha^{-1}, 24.86–55.74 t C ha^{-1}, and 36.24–61.69 t C ha^{-1} for the whole study area, Hangjiahu plain (N), Jinqu basin (B), and coastal regions (C1, C2, and C3), respectively. The areas rendered with white color in (a) are non-farmlands.

SOC estimates using a random forest model. Although kriging interpolation methods have been widely used in estimating unsampled locations from observations, some uncertainties have emerged as a result of its over smoothing in spatial interpolation (Olea and Pawlowsky, 1996). In this study, the data-induced uncertainties (i.e., 95% CI) associated with the SOC stock prediction were estimated to be 35.40–63.21 t C ha^{-1}, 35.54–66.40 t C ha^{-1}, 24.86–55.74 t C ha^{-1}, and 36.24–61.69 t C ha^{-1} for the whole study area, Hangjiahu plain, Jinqu basin, and coastal plain, respectively (Fig. 4a). The majority of the standardized 95% CIs were lower than 1.0 and followed a normal distribution with μ = 0.55 (Fig. 4b). Replacing the over-smoothed results by the geostatistical simulation estimates can partially capture the prediction uncertainties that emerged from spatial-interpolated predictor inputs. Certainly, some uncertainties of inputs that are unable to be quantified remain. Specifically, the data quality (e.g., sampling representativeness and accuracy) from the field survey and laboratory analysis could not be quantified. For instance, the original triplicate measurements of SOC were inaccessible at different testing institutions. Additionally, the uncertainties of satellite retrievals (e.g., land cover) were also difficult to determine.

The model estimation uncertainties, measured by coefficients of variation (CV), were generally lower than 0.2. Relatively higher CV (> 0.3) were found in small areas scattered throughout the central and southern areas (Figs. b and S5). The estimation uncertainties of the random forest in this study were lower than those of national SOC estimates in Australia, where the standardized 95% CI ranged from 0.1 to 2.7, with a CV of 0.03–0.69 (Viscarra Rossel et al., 2014). While the estimation uncertainties of the random forest model were largely reduced by optimizing the model hyper parameters and refining the predictor variables (Figs. S4 and S6), a considerable amount of uncertainty remained in the estimation due to the limited representativeness of the training data. The prediction nodes beyond the range of training data (i.e., extrapolation issues), are expected to suffer from high uncertainty (Horning, 2010). Therefore, large training datasets are always favored for achieving reliable estimates. In this study, we surveyed nearly 30,000 sampling sites which comprehensively spanned relevant features including soil properties, farmland topographic conditions, land use, and cropping patterns. The representativeness and reliability of the resulting model prediction undoubtedly benefited from this large dataset.

The random forest model is always subject to several limitations, including interpretability, data requirements, and the availability of informative variables. While variable importance measures and partial dependence plots are valuable methods for model interpretation, the model structure of random forest is rather complicated, causing interpretability that is less clear than classic statistical models such as GLM (Dormann et al., 2013). Since the main purpose of this study is to obtain accurate SOC estimates, prediction performance outweighed interpretability. In addition, the random forest model generally requires a large training dataset, which limits its application to small-sample studies, since collecting soil samples and measuring properties are laborious and costly. Furthermore, as a data-driven approach, the random forest model relies highly on the availability and effectiveness of input predictor variables. Informative predictors, such as fertilizer application rate in this study, improve the performance in predicting SOC stock (Fig. 5), but the redundant variables tend to degrade the model

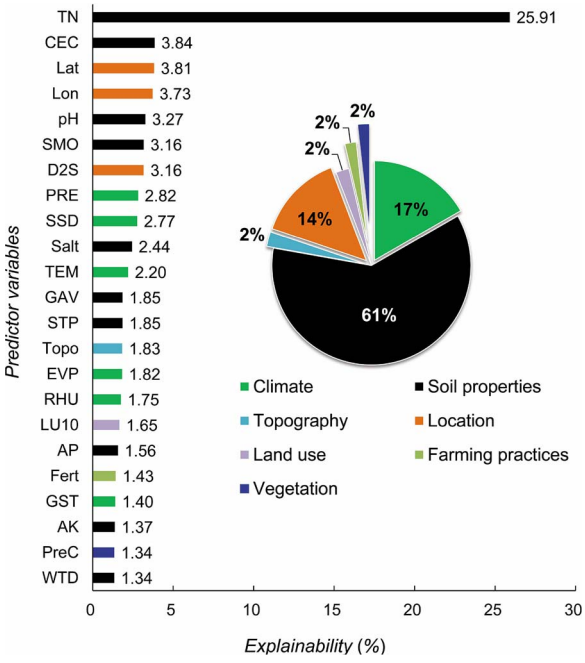

Fig. 5. Relative importance of the predictor variables in the random forest model. The importance of each variable was determined by the total decrease in node impurities from splitting on that predictor variable averaged over all trees. The categorical importance for predictor variables were summarized to 100%. The variable categorization and the variable abbreviations are listed in Table 1 and Table S2, respectively.

performance (Fig. S4). Through the variable selection, the redundant variables were excluded from the random forest model for simplifying model structure and improving prediction performance. In the future, more feature engineering work for exploring potentially informative predictors, e.g., the Landsat bands revealed by Schillaci et al. (2017), should be conducted to further reduce the estimation uncertainty.

3.5. Relationships of SOC with soil properties, climate, and satellite retrievals

The variable importance results show that soil properties have the most important effects on predicting the SOC stocks, accounting for 61% of the total relative importance. The most important variable for SOC dynamics was the soil total nitrogen content (TN) with an importance value of 26% (Fig. 5), which is consistent with results found in previous studies (Elbasiouny et al., 2014; Were et al., 2015). A nearly linear relationship between SOC and soil TN in the partial dependence analysis also indicated this strong association (Fig. S7). The strong relationship between the soil TN and SOC probably originated from the tightly-coupled cycling of carbon and nitrogen in the plant-soil agroecosystem (Were et al., 2015). A common feature of this cycle is that the microorganisms tend to utilize a fixed C/N ratio (usually 30–35:1) (Prahl et al., 1994). Also, the nitrogen transformation processes common in soils, e.g., nitrification and denitrification, use the energy from carbon (Batlle-Aguilar et al., 2011). Other soil properties, such as pH, CEC and moisture, also play important roles in the spatial distributions of SOC stock (Figs. and S7). These various chemical properties of rhizosphere soils are likely to favor/restrain SOC decomposition/turnover by stimulating/inhibiting the activities of soil microbes (You et al., 1999; Kemmitt et al., 2006). Additionally, different soil types contain different soil aggregate compositions. For instance, the stable microaggregates in Anthrosols (usually paddy field) protect the SOC from being decomposed by microorganisms (Wissing et al., 2013; Wang et al., 2014). However, due to the absence of extensive data for these variables, few studies have addressed these soil properties for SOC prediction. Hence, the accurate and spatially comprehensive data from soil properties should be included to improve the SOC prediction.

Aside from the importance of soil properties, climatic conditions, fertilization and satellite retrievals explain additional sources of variation in SOC estimates, with summed importance values of 14%, 2% and 2%, respectively (Fig. 5). Precipitation is the most important climatic condition, followed by sunshine duration, temperature, evaporation, relative humidity and ground surface temperature. Climatic conditions generally dominate the control of plant growth (i.e., biomass production) and biological activities, which in turn contribute to different levels of SOC accumulation through affecting residue decomposition in shallow soil layers (Jobbagy and Jackson, 2000). The effects of climatic conditions on SOC stock have proven to be dominant factors at national, continental, or global levels (Viscarra Rossel and Chen, 2011). A moderate importance of climatic conditions on SOC stock is revealed in this regional scale study. Nitrogenous fertilizer application generally increases the crop biomass and various strategies of returning plant residues to soil also play important roles in the SOC formation. For instance, high fertilizer application rates occurred in the vegetable fields but the majority of the crop biomass was harvested as food with little residues returned to soils, resulting in a low contribution to SOC stock (Wang and Zhou, 2011). The satellite retrieval, e.g., land use, was not a strong predictor in this study (Fig. 5). Moreover, both the NDVI and elevation were excluded from the final model, which differed from their important roles in predicting SOC stock across forests, grasslands, and cropland in some previous studies (Viscarra Rossel and Chen, 2011; Akpa et al., 2016). This difference is probably due to the relatively minor spatial variability of these factors for Zhejiang farmlands, and thus the importance of these variables on SOC predictions decreased. Since satellite retrievals have spatially continuous features and universal availability, they might also act as hidden predictors for future SOC estimations on large scales when potentially effective indicators could be found (Viscarra Rossel and Chen, 2011).

3.6. Potential discrepancies between this study and HWSD

Our results show that the spatial SOC stock might have changed considerably in the last two decades. The overall difference in the SOC baselines between HWSD and this study is estimated to be 10.4 t C ha^{-1} (Table 2). At sub-regional levels, large differences were observed in the valley terrace and Jinqu basin, where the average SOC stock differences were up to 20 t C ha^{-1}. In the coastal plain, the average SOC stock increased by approximately 10 t C ha^{-1} (Table 2). In the Hangjiahu plain, the SOC stock was relatively stable (Table 2; Figs. a and S9). The comparison results suggest that carbon losses might take place across Zhejiang province from early 1980s to 2008, especially in the central Hangjiahu plain, Jinqu basin, and the scattered valley terrace (Fig. S10). In contrast, in most parts of the coastal regions and the outage of Hangjiahu plain, the farmland soil might sequester a large amount of carbon in recent decades (Figs. a and S9).

Changes of farming practices (e.g., tillage practices, farmland levelling, and cropping system shifts) in the past two decades tended to drive the SOC differences in the farmlands. For instance, the dominant cropping system in early 1980s was double paddy rice in the Jinqu basin, Hangjiahu plain, and the scattered valley terrace, but this has gradually shifted to dry-land crops (e.g., vegetables, fruits, maize, and broad bean) or single paddy rice rotations with these dry-land crops (ZJBS, 2016). The root residues returned to the soil from these dry crops are usually lower than those in the paddy rice cropping systems. Furthermore, more frequent tillage practices in these dry-land farmlands directly disrupt the stability of the physical structure of topsoil. For example, the SOC decomposition in dry lands would be accelerated compared with those in paddy fields due to increased aeration porosity and weaker soil aggregate stability in dry land soil (Six et al., 2000; Wissing et al., 2013). Accordingly, an opposite shift of the cropping system, i.e., from mulberry (a dry-land crop) to vegetable-paddy rice rotations in the outage of Hangjiahu plain, likely benefited the soil carbon sequestration. During recent years, aquaculture-vegetables rotations and farmland levelling have been employed for many farmlands in the central Hangjiahu plain and most of Zhejiang, respectively, and consequently the topsoil was completely removed and SOC dramatically changed (Fig. S10). In addition, the increase of SOC stock in the coastal areas was probably caused by a reclaimed soil ripening that was accompanied by the intensive application of compound fertilizers and organic fertilizers (Wang and Zhou, 2011; ZJBS, 2016). These management practices would have increased the organic carbon accumulation by increasing the soil carbon input, as well as improving soil physicochemical properties (e.g., desalination and dealkalization) and the stability of macroaggregate structures (Wissing et al., 2013; Li et al., 2014).

While these shifts in farming practices were likely responsible for the differences of these two SOC baselines, the inconsistencies of sample sizes and SOC mapping methods tended to contribute to the uncertainties associated with the comparisons. The HWSD database was derived from relatively sparse data, with only 7292 soil samples for all of China and it was partially derived from the World Inventory of Soil Emission Potential (WISE). Less than 490 samples were located in Zhejiang (ZJSSO, 1994; FAO/IIASA/ISRIC/ISSCAS/JRC, 2012), far less than the extensive soil samples (29,927) used in this study. The large difference in the sampling data sizes could lead to different estimation accuracies, particularly at fine resolutions, with small sample sizes generally leading to higher mapping uncertainties (Wisz et al., 2008). In addition, the SOC content in the HWSD baseline was estimated by extending the sparse soil profiles according to soil map units (FAO/IIASA/ISRIC/ISSCAS/JRC, 2012). The soil map units were usually much coarser than the 1 km resolution adopted by HWSD, and tended to be inconsistent with the spatial units of soil management practices,

climate, and land use changes (Smith et al., 2012). Consequently, the SOC predictions were inaccurate for those fine cells with similar soil management practice/land use characteristics but located in different soil map units. The standard deviations of SOC in our estimates were much lower than those of HWSD (Table 2). While high uncertainties emerged in fine scale comparisons, the spatial patterns of the mapped SOC stock were generally consistent with the spatial distributions of the measured SOC levels at large scales (Figs. 1 and 2a). Given the current data availability, the comparison of these two SOC baselines provided valuable information on farmland SOC stock changes in Zhejiang province.

3.7. Implications for carbon accounting

The SOC baseline map generated in this study is critical for obtaining up-to-date carbon accounting for agricultural soil at a regional level. A straightforward application of the SOC baseline map is to obtain reliable estimates of carbon sequestration/loss for agricultural soils by using the stock-difference method, and this is one of the carbon accounting methods proposed by IPCC (Bird et al., 2010). Compared with the SOC stock estimated from HWSD, the estimate generated by this study was approximately 17% lower for Zhejiang province (Table 2). Extrapolating this over-estimation ratio to the whole East China, 239 Mt C (1 Mt = 10^6 t) might be over-estimated in the HWSD in total (NASMG, 2016). Two recent studies have also corroborated the spatiotemporal changes between the previously estimated and recent SOC baselines in both Australia and China (Liao et al., 2009; Viscarra Rossel et al., 2014). Hence, this updated regional SOC stock is favored for reliable carbon accounting in agricultural soils, especially for regions with frequent cropping rotations and intensive tillage practices in recent decades.

The updated SOC baseline at a fine spatial resolution serves to improve the accuracy of the carbon accounting by supporting the more spatially resolved inputs for process-based simulation models. The initial SOC content is used as an essential input in regional environment modeling (e.g., DNDC, CENTURY) for carbon accounting in the IPCC Tier 3 method under the impacts of land use dynamics, soil management, and climate change (IPCC, 2006; Lugato et al., 2014). However, the current SOC content inputs for carbon accounting at regional levels are usually derived from the soil inventories (e.g., HWSD, WISE) surveyed decades ago (FAO/IIASA/ISRIC/ISSCAS/JRC, 2012). Therefore, the new updated SOC baseline can mitigate the error propagation that has emerged in subsequent simulations (e.g. N_2O/CO_2 emissions). In addition, the finer spatial resolution of the current assessment also allows the running of various environmental models with higher spatial precision. Our newly predicted SOC baseline is in accordance with an increased demand for the environmental modeling of carbon accounting (Romero et al., 2012), and can be used for quantifying greenhouse gas emissions from farmlands.

The spatially resolved estimates of SOC baseline could also guide soil monitoring for carbon accounting in the future. On the one hand, an updated sampling strategy should collect more samples from the areas with high estimation uncertainties (Fig. 2b). On the other hand, the spatially detailed discrepancies can help identify the potential farmland soil for carbon sequestration, particularly where the farmlands either underwent large amounts carbon loss (Fig. S10) or contained low SOC stock (Fig. 2a). The management practices for areas with high SOC stock should be promoted for improving the larger regional SOC stock. Effective soil management practices, such as the organic fertilizer applications, cropping system shifts from dryland cropping to the paddy rice, and conservational tillage (Lal, 2008; Deng et al., 2016), can be more widely adopted to capture carbon and thereby mitigate the carbon emissions from those agricultural soils.

4. Conclusion

An SOC baseline map at a spatial resolution of 250 m for the farmlands of Zhejiang province, East China was developed by using the random forest model. With a comprehensive set of predictor variables, the random forest model exhibited a high predictive performance with R^2 of 0.76. The model prediction showed that this new estimate of SOC density (50.1 t C ha^{-1}) was much lower than SOC in HWSD (60.5 t C ha^{-1}) which was accompanied by high spatial discrepancies. Extrapolating these discrepancies to the other regions with intensive anthropogenic activities, considerable amounts of soil carbon estimation biases and uncertainties tended to emerge in carbon accounting and modeling. Hence, for Zhejiang province and regions with similar cultivation practices, it is highly recommended to use new the SOC baselines for regional farmland management, carbon accounting, and biogeochemical simulations. As the main data source for carbon accounting under climate change, the data quality of HWSD should be re-evaluated all over the world, especially for intensively cultivated regions. Various programs for updating and improving soil property databases have been proposed and/or implemented, such as the Land Use/Cover statistical Area frame Survey for Europe (EUROSTAT, 2016) and a national soil survey for China (MOA, 2014), and these would provide more recent large-scale SOC data to guide scientific research and management efforts.

Acknowledgements

We are grateful to Hong Lu, Fei Wang, Fangjin Qin, Yiding Chen, Tangxiao Cen, Wuying Bian, and other colleagues for their assistance in field sampling, laboratory analyses, and document collecting. We also thank the anonymous reviewers for their valuable comments. This study was supported by the National Key Research and Development Program of China (No. 2016YFD0200802).

Appendix A. Supplementary data

SOC stock estimation methods, the variable correlation matrix, partial dependence results, and spatial distributions of SOC in HWSD, can be found in the online version. The SOC baseline map for Zhejiang developed in this study can be obtained by contacting the corresponding author, Yu Zhan.

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.agee.2017.11.022.

References

Akpa, S.I.C., Odeh, I.O.A., Bishop, T.F.A., Hartemink, A.E., Amapu, I.Y., 2016. Total soil organic carbon and carbon sequestration potential in Nigeria. Geoderma 271, 202–215.
Batjes, N.H., 2009. Harmonized soil profile data for applications at global and continental scales: updates to the WISE database. Soil Use Manage. 25, 124–127.
Batlle-Aguilar, J., Brovelli, A., Porporato, A., Barry, D.A., 2011. Modelling soil carbon and nitrogen cycles during land use change. A review. Agron. Sustain. Dev. 31, 251–274.
Bell, M.J., Worrall, F., 2009. Estimating a region's soil organic carbon baseline: the undervalued role of land-management. Geoderma 152, 74–84.
Bird, D.N., Pena, N., Schwaiger, H., Zanchi, G., 2010. Review of Existing Methods for Carbon Accounting. Review of Existing Methods for Carbon Accounting. Occasional Paper 54. CIFOR, Bogor, Indonesia.
Breiman, L., 2001. Random forests. Mach Learn 45, 5–32.
CMDC (China Meteorological Data Center), 2015. Daily-value Datasets of Ground Surface in China V3.0. http://data.cma.cn/data/detail/dataCode/SURF_CLI_CHN_MUL_DAY_V3.0.html. (Accessed 9 June 2015).
Chaplot, V., Bouahom, B., Valentin, C., 2010. Soil organic carbon stocks in Laos: spatial variations and controlling factors. Glob. Change Biol. 16, 1380–1393.
Cutler, D.R., Edwards, T.C., Beard, K.H., Cutler, A., Hess, K.T., Gibson, J., Lawler, J.J., 2007. Random forests for classification in ecology. Ecology 88, 2783–2792.
Deng, X., Zhan, Y., Wang, F., Ma, W., Ren, Z., Chen, X., Qin, F., Long, W., Zhu, Z., Lv, X., 2016. Soil organic carbon of an intensively reclaimed region in China: current status and carbon sequestration potential. Sci. Total Environ. 565, 539–546.
Deutsch, C.V., Journel, A.G., 1998. GSLIB: Geostatistical Software Library and User's Guide, second edition. Oxford University Press, New York, Oxford.

Dormann, C.F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., Marquéz, J.R.G., Gruber, B., Lafourcade, B., Leitão, P.J., Münkemüller, T., McClean, C., Osborne, P.E., Reineking, B., Schröder, B., Skidmore, A.K., Zurell, D., Lautenbach, S., 2013. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. Ecography 36, 27–46.

EUROSTAT, 2016. Land Use and Land Cover Area Frame Survey. http://ec.europa.eu/eurostat/web/lucas/overview. (Accessed 9 November 2016).

Elbasiouny, H., Abowaly, M., Abu_Alkheir, A., Gad, A.A., 2014. Spatial variation of soil carbon and nitrogen pools by using ordinary Kriging method in an area of north Nile Delta, Egypt. Catena 113, 70–78.

Elith, J., Leathwick, J.R., Hastie, T., 2008. A working guide to boosted regression trees. J. Anim. Ecol. 77, 802–813.

FAO/IIASA/ISRIC/ISSCAS/JRC, 2012. Harmonized World Soil Database (version 1.2). FAO, Rome, Italy and IIASA, Laxenburg, Austria.

Fantappiè, M., L'Abate, A., Costantini, E.A.C., 2010. Factors influencing soil organic carbon stock variations in Italy during the last three decades. In: Zdruli, Pagliai, Kapur, Land, F. (Eds.), Land Degradation and Desertification: Assessment, Mitigation and Remediation. Springer, Dordrecht Heidelberg London New York.

Farina, R., Marchetti, A., Francaviglia, R., Napoli, R., Bene, C.D., 2017. Modeling regional soil C stocks and CO_2 emissions under Mediterranean cropping systems and soil types. Agric. Ecosyst. Environ. 238, 128–141.

Grimm, R., Behrens, T., Märker, M., Elsenbeer, H., 2008. Soil organic carbon concentrations and stocks on Barro Colorado Island — digital soil mapping using Random Forests analysis. Geoderma 146, 102–113.

Gromski, P.S., Correa, E., Vaughan, A.A., Wedge, D.C., Turner, M.L., Goodacre, R., 2014. A comparison of different chemometrics approaches for the robust classification of electronic nose data. Anal. Bioanal. Chem. 406, 7581–7590.

Halvorson, A.D., Wienholdb, B.J., Black, A.L., 2002. Tillage, nitrogen, and cropping system effects on soil carbon sequestration. Soil Sci. Soc. Am. J. 66, 906–912.

Hengl, T., Heuvelink, G.B., Kempen, B., Leenaars, J.G., Walsh, M.G., Shepherd, K.D., Sila, A., MacMillan, R.A., Mendes de Jesus, J., Tamene, L., Tondoh, J.E., 2015. Mapping soil properties of africa at 250 m resolution: random forests significantly improve current predictions. PLoS One 10, e0125814.

Hengl, T., Mendes de Jesus, J., Heuvelink, G.B., Ruiperez Gonzalez, M., Kilibarda, M., Blagotic, A., Shangguan, W., Wright, M.N., Geng, X., Bauer-Marschallinger, B., Guevara, M.A., Vargas, R., MacMillan, R.A., Batjes, N.H., Leenaars, J.G., Ribeiro, E., Wheeler, I., Mantel, S., Kempen, B., 2017. SoilGrids250m: global gridded soil information based on machine learning. PLoS One 12, e0169748.

Hiemstra, P.H., Pebesma, E.J., Twenhöfel, C.J.W., Heuvelink, G.B.M., 2009. Real-time automatic interpolation of ambient gamma dose rates from the Dutch radioactivity monitoring network. Comput. Geosci.–U. K. 35, 1711–1721.

Holben, B.N., 1986. Characteristics of maximum-value composite images from temporal AVHRR data. Int. J. Remote Sens. 7, 1417–1434.

Horning, N., 2010. Random Forests: an algorithm for image classification and generation of continuous fields data sets. In: Proceedings of the International Conference on Geoinformatics for Spatial Infrastructure Development in Earth and Allied Sciences. Osaka, Japan.

IPCC, 2006. Cropland. In: In: Eggleston, H.S., Buendia, L., Miwa, K., Ngara, T., Tanabe, K. (Eds.), 2006 IPCC Guidelines for National Greenhouse Gas Inventories, vol. IV. pp. 1–66.

Jobbagy, E.g., Jackson, R.b., 2000. The vertical distribution of soil organic carbon and its relation to climate and vegetation. Ecol. Appl. 10, 423–436.

Kemmitt, S.J., Wright, D., Goulding, K.W.T., Jones, D.L., 2006. pH regulation of carbon and nitrogen dynamics in two agricultural soils. Soil Biol. Biochem. 38, 898–911.

Lal, R., 2008. Sequestration of atmospheric CO2 in global carbon pools. Energy Environ. Sci. 1, 86.

Li, J., Heap, A.D., Potter, A., Daniell, J.J., 2011. Application of machine learning methods to spatial interpolation of environmental variables. Environ. Modell. Softw. 26, 1647–1659.

Li, J., Pu, L., Zhu, M., Zhang, J., Li, P., Dai, X., Xu, Y., Liu, L., 2014. Evolution of soil properties following reclamation in coastal areas: a review. Geoderma 226–227, 130–139.

Liao, Q., Zhang, X., Li, Z., Pan, G., Smith, P., Jin, Y., Wu, X., 2009. Increase in soil organic carbon stock over the last two decades in China's Jiangsu Province. Glob. Change Biol. 15, 861–875.

Liaw, A., Wiener, M., 2002. Classification and Regression by RandomForest, vol. 2. R News, pp. 18–22.

Liu, D., Wang, Z., Zhang, B., Song, K., Li, X., Li, J., Li, F., Duan, H., 2006. Spatial distribution of soil organic carbon and analysis of related factors in croplands of the black soil region, Northeast China. Agric. Ecosyst. Environ. 113, 73–81.

Lugato, E., Panagos, P., Bampa, F., Jones, A., Montanarella, L., 2014. A new baseline of organic carbon stock in European agricultural soils using a modelling approach. Glob. Change Biol. 20, 313–326.

MLR, Ministry of Land and Resources of the People's Republic of China, 2014. Bulletin of Important Data Product from Second Land Survey of Zhejiang Province. http://www.mlr.gov.cn/tdzt/tdgl/decdc/dccg/gscg/201407/P020140724562862286636.pdf. (Accessed 9 December 2014).

MOA, Ministry of Agriculture of the People's Republic of China, 2006. Soil Testing Part6: Method for Determination of Soil Organic Matter, NY/T1121.6. (in Chinese).

MOA, Ministry of Agriculture of the People's Republic of China, 2014. Bulletin of Grade of Cultivation Land from National Survey. http://www.moa.gov.cn/zwllm/zwdt/201412/t20141217_4298677.htm. (Accessed 17 May 2017).

Ma, K., Liu, J., Balkovič, J., Skalský, R., Azevedo, L.B., Kraxner, F., 2016. Changes in soil organic carbon stocks of wetlands on China's Zoige plateau from 1980 to 2010. Ecol. Model. 327, 18–28.

Martin, D., Lal, T., Sachdev, C.B., Sharma, J.P., 2010. Soil organic carbon storage changes with climate change, landform and land use conditions in Garhwal hills of the Indian Himalayan mountains. Agric. Ecosyst. Environ. 138, 64–73.

Martin, M.P., Orton, T.G., Lacarce, E., Meersmans, J., Saby, N.P.A., Paroissien, J.B., Jolivet, C., Boulonne, L., Arrouays, D., 2014. Evaluation of modelling approaches for predicting the spatial distribution of soil organic carbon stocks at the national scale. Geoderma 223–225, 97–107.

Meersmans, J., Martin, M.P., De Ridder, F., Lacarce, E., Wetterlind, J., De Baets, S., Le Bas, C., Louis, B.P., Orton, T.G., Bispo, A., Arrouays, D., 2012. A novel soil organic C model using climate, soil type and management data at the national scale in France. Agron. Sustain. Dev. 32, 873–888.

Miller, A.J., Amundson, R., Burke, I.C., Yonker, C., 2004. The effect of climate and cultivation on soil organic C and N. Biogeochemistry 67, 57–72.

Mishra, U., Lal, R., Liu, D., Van Meirvenne, M., 2010. Predicting the spatial variation of the soil organic carbon pool at a regional scale. Soil Sci. Soc. Am. J. 74, 906.

Muñoz-Rojas, M., Doro, L., Ledda, L., Francaviglia, R., 2015. Application of CarboSOIL model to predict the effects of climate change on soil organic carbon stocks in agro-silvo-pastoral Mediterranean management systems. Agric. Ecosyst. Environ. 202, 8–16.

NASALPDAAC, 2015. SRTMGL1: NASA Shuttle Radar Topography Mission Global 1 Arc Second V003. NASA EOSDIS Land Processes DAAC. USGS Earth Resources Observation and Science (EROS) Center, Sioux Falls, South Dakota(https://lpdaac.usgs.gov). https://10.5067/MEaSUREs/SRTM/SRTMGL1.003. (Accessed 27 October 2015).

NASA LP DAAC, 2016. MOD13Q1 MODIS/Terra Vegetation Indices 16-Day L3 Global 250 M Grid SIN V006. NASA EOSDIS Land Processes DAAC, USGS Earth Resources Observation and Science (EROS) Center, Sioux Falls, South Dakota(https://lpdaac.usgs.gov). https://doi.org/10.5067/MODIS/MOD13Q1.006. (Accessed 18 August 2016).

NASMG, 2016. National Administrator of Surveying Mapping and Geoinformation of China. Global Land Cover 30 m (GLC2010). http://www.globallandcover.org/GLC30Download/index.aspx. (Accessed 21 March 2016).

Ni, J., 2001. Carbon storage in terrestrial ecosystems of China: estimates at different spatial resolutions and their responses to climate change. Clim. Change 49, 339–358.

Olea, R.A., Pawlowsky, V., 1996. Compensating for estimation smoothing in kriging. Math. Geol. 28, 407–417.

Pan, G., Xu, X., Smith, P., Pan, W., Lal, R., 2010. An increase in topsoil SOC stock of China's croplands between 1985 and 2006 revealed by soil monitoring. Agric. Ecosyst. Environ. 136, 133–138.

Perlman, J., Hijmans, R.J., Horwath, W.R., 2014. A metamodelling approach to estimate global N2O emissions from agricultural soils. Glob. Ecol. Biogeogr. 23, 912–924.

Prahl, F.G., Ertel, J.R., Goni, M.A., Sparrow, M.A., Eversmeyer, B., 1994. Terrestrial organic carbon contributions to sediments on the Washington margin. Geochim. Cosmochim. Acta 58, 3035–3048.

R Core Team, 2016. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria ISBN 3-900051-07-0, URL. http://www.R-project.org/.

Reid, C.E., Jerrett, M., Petersen, M.L., Pfister, G.G., Morefield, P.E., Tager, I.B., Raffuse, S.M., Balmes, J.R., 2015. Spatiotemporal prediction of fine particulate matter during the 2008 northern California wildfires using machine learning. Environ. Sci. Technol. 49, 3887–3896.

Romero, C.C., Hoogenboom, G., Baigorria, G.A., Koo, J., Gijsman, A.J., Wood, S., 2012. Reanalysis of a global soil database for crop and environmental modeling. Environ. Modell. Softw. 35, 163–170.

Schillaci, C., Lombardo, L., Saia, S., Fantappiè, M., Märker, M., Acutis, M., 2017. Modelling the topsoil carbon stock of agricultural lands with the Stochastic Gradient Treeboost in a semi-arid Mediterranean region. Geoderma 286, 35–45.

Six, J., Elliott, E.T., Paustian, K., 2000. Soil macroaggregate turnover and microaggregate formation: a mechanism for C sequestration under no-tillage agriculture. Soil Biol. Biochem. 32, 2099–2103.

Smith, P., Davies, C.A., Ogle, S., Zanchi, G., Bellarby, J., Bird, N., Boddey, R.M., McNamara, N.P., Powlson, D., Cowie, A., Noordwijk, M., Davis, S.C., Richter, D.D.E.B., Kryzanowski, L., Wijk, M.T., Stuart, J., Kirton, A., Eggar, D., Newton-Cross, G., Adhya, T.K., Braimoh, A.K., 2012. Towards an integrated global framework to assess the impacts of land use and management change on soil carbon: current capability and future vision. Glob. Change Biol. 18, 2089–2101.

Stoorvogel, J.J., Bakkenes, M., Temme, A.J.A.M., Batjes, N.H., ten Brink, B.J.E., 2017. S-world: a global soil map for environmental modelling. Land Degrad. Dev. 28, 22–33.

Tang, H., Qiu, J., Wang, L., Li, H., Li, C., van Ranst, E., 2010. Modeling soil organic carbon storage and its dynamics in croplands of China. Agric. Sci. China 9, 704–712.

Villarino, S.H., Studdert, G.A., Laterra, P., Cendoya, M.G., 2014. Agricultural impact on soil organic carbon content: testing the IPCC carbon accounting method for evaluations at county scale. Agric. Ecosyst. Environ. 185, 118–132.

Viscarra Rossel, R.A., Chen, C., 2011. Digitally mapping the information content of visible–near infrared spectra of surficial Australian soils. Remote Sens. Environ. 115, 1443–1455.

Viscarra Rossel, R.A., Webster, R., Bui, E.N., Baldock, J.A., 2014. Baseline map of organic carbon in Australian soil to support national carbon accounting and monitoring under climate change. Glob. Chang Biol. 20, 2953–2970.

Wager, S., Hastie, T., Efron, B., 2014. Confidence intervals for random forests: the jackknife and the infinitesimal jackknife. J. Mach. Learn. Res. 15, 1625–1651.

Wang, F., Zhou, Z., 2011. Evaluation of Soil Fertility in Arable Land and Its Improvement in Ningbo City (in Chinese). Zhejiang University Press, Hangzhou.

Wang, H., Guan, D., Zhang, R., Chen, Y., Hu, Y., Xiao, L., 2014. Soil aggregates and organic carbon affected by the land use change from rice paddy to vegetable field. Ecol. Eng. 70, 206–211.

Were, K., Bui, D.T., Dick, Ø.B., Singh, B.R., 2015. A comparative assessment of support

vector regression, artificial neural networks, and random forests for predicting and mapping soil organic carbon stocks across an Afromontane landscape. Ecol. Indic. 52, 394–403.
West, T.O., Post, W.M., 2002. Soil organic carbon sequestration rates by tillage and crop rotation: a global data analysis. Soil Sci. Soc. Am. J. 66, 1930–1946.
Wissing, L., Kölbl, A., Häusler, W., Schad, P., Cao, Z.-H., Kögel-Knabner, I., 2013. Management-induced organic carbon accumulation in paddy soils: the role of organo-mineral associations. Soil Till. Res. 126, 60–71.
Wisz, M.S., Hijmans, R.J., Li, J., Peterson, A.T., Graham, C.H., Guisan, A., 2008. Effects of sample size on the performance of species distribution models. Divers. Distrib. 14, 763–773.
Xiong, X., Grunwald, S., Myers, D.B., Ross, C.W., Harris, W.G., Comerford, N.B., 2014. Interaction effects of climate and land use/land cover change on soil organic carbon sequestration. Sci. Total Environ. 493, 974–982.
Yan, Y., Tian, J., Fan, M., Zhang, F., Li, X., Christie, P., Chen, H., Lee, J., Kuzyakov, Y., Six, J., 2012. Soil organic carbon and total nitrogen in intensively managed arable soils. Agric. Ecosyst. Environ. 150, 102–110.
You, S.-J., Yin, Y., Allen, H.E., 1999. Partitioning of organic matter in soils: effects of pH and water/soil ratio. Sci. Total Environ. 227, 155–160.
Yu, Y., Huang, Y., Zhang, W., 2012. Modeling soil organic carbon change in croplands of China, 1980–2009. Glob. Planet. Change 82–83, 115–128.
ZASMG, Zhejiang Administration of Surveying Mapping and Geoinformation, 2013. Digital Atlas of Land Use in Zhejiang. http://183.129.171.247:18089/page/serve/stats.html. (Accessed 22 September 2013).
ZJBS, Zhejiang Provincial Bureau of Statistics, 2016. Statistical Yearbook of Zhejiang Province. http://www.zj.stats.gov.cn/tjsj/tjnj/. (Accessed 11 April 2016).
ZJSSO, Zhejiang Soil Survey Office, 1994. Zhejiang Soil (in Chinese). Zhejiang Science and Technology Press, Hanghou.
Zhang, H., Wu, P., Yin, A., Yang, X., Zhang, M., Gao, C., 2017. Prediction of soil organic carbon in an intensively managed reclamation zone of eastern China: a comparison of multiple linear regressions and the random forest model. Sci. Total Environ 592, 704–713.